SIAM J. MATRIX ANAL. APPL.
Vol. 30, No. 1, pp. 121–141

GRAPH CLUSTERING VIA A DISCRETE UNCOUPLING PROCESS*

STIJN VAN DONGEN†

Abstract. A discrete uncoupling process for finite spaces is introduced, called the *Markov Cluster Process* or the *MCL process*. The process is the engine for the graph clustering algorithm called the *MCL algorithm*. The *MCL process* takes a stochastic matrix as input, and then alternates expansion and inflation, each step defining a stochastic matrix in terms of the previous one. Expansion corresponds with taking the kth power of a stochastic matrix, where $k \in \mathbb{N}$. Inflation corresponds with a parametrized operator Γ_r, $r \geq 0$, that maps the set of (column) stochastic matrices onto itself. The image $\Gamma_r M$ is obtained by raising each entry in M to the rth power and rescaling each column to have sum 1 again. In practice the process converges very fast towards a limit that is invariant under both matrix multiplication and inflation, with quadratic convergence around the limit points. The heuristic behind the process is its expected behavior for (Markov) graphs possessing cluster structure. The process is typically applied to the matrix of random walks on a given graph G, and the connected components of (the graph associated with) the process limit generically allow a clustering interpretation of G. The limit is in general extremely sparse and iterands are sparse in a weighted sense, implying that the *MCL algorithm* is very fast and highly scalable. Several mathematical properties of the *MCL process* are established. Most notably, the process (and algorithm) iterands posses structural properties generalizing the mapping from process limits onto clusterings. The inflation operator Γ_r maps the class of matrices that are diagonally similar to a symmetric matrix onto itself. The phrase *diagonally positive semi-definite* (*dpsd*) is used for matrices that are diagonally similar to a positive semi-definite matrix. For $r \in \mathbb{N}$ and for M a stochastic *dpsd* matrix, the image $\Gamma_r M$ is again *dpsd*. Determinantal inequalities satisfied by a *dpsd* matrix M imply a natural ordering among the diagonal elements of M, generalizing the mapping of process limits onto clusterings. The spectrum of $\Gamma_\infty M$ is of the form $\{0^{n-k}, 1^k\}$, where k is the number of endclasses of the ordering associated with M, and n is the dimension of M. This attests to the uncoupling effect of the inflation operator.

Key words. stochastic uncoupling, graph clustering, Markov graph, Markov matrix, diagonal similarity, positive semi-definite matrices, circulant matrices

AMS subject classifications. 68R10, 05C85, 05C90

DOI. 10.1137/040608635

1. Introduction.

The subject of study is a parametrized algebraic process called the Markov Cluster Process (*MCL* process), which is the engine of a cluster algorithm for graphs, accordingly named the *MCL* algorithm. The algorithm is nothing more than a shell in which parameters are set, the *MCL* process is computed, and the result is interpreted. The process itself is defined on the space of stochastic matrices. Given a graph G, the algorithm employs the process by applying it to the matrix of random walks on G.

The *MCL* algorithm [11, 12] was first applied in the field of protein family detection [18]. In this setting, proteins are nodes in a graph where the edge weights are derived from BLAST (Basic Local Alignment Search Tool) scores between protein amino-acid sequences. Following [18], the algorithm has been widely applied in bioinformatics, in a diversity of settings and applications.

*Received by the editors May 19, 2004; accepted for publication (in revised form) by D. A. Bini July 31, 2007; published electronically February 20, 2008.
http://www.siam.org/journals/simax/30-1/60863.html

†Wellcome Trust Genome Campus, The Wellcome Trust Sanger Institute, Hinxton CB10 1SA Cambridge, United Kingdom (svd@sanger.ac.uk). This author's research was carried out at the CWI, the National Research Institute for Mathematics and Computer Science in the Netherlands.

A number of publications have used *MCL* for large scale single species or cross-species protein and gene family analysis, e.g., [15, 16, 28, 38, 52, 60]. Other protein-related *MCL* applications in bioinformatics are large scale sequence space analysis [19, 37], hybrid *MCL*/single-link clustering [29], orthologous groups [41], kinase proteins [24], secreted proteins [7], eye proteins [39], mobile genetic elements [40], protein interaction networks [5, 54], and protein function determination [61]. Additionally, *MCL* has been applied in corpus linguistics [13, 14, 25], content-based image retrieval [32], peer-to-peer network analysis [57], and social network analysis [46].

Factors aiding the adoption of the *MCL* algorithm include (a) It generates well-balanced flat (nonhierarchic) clusterings. (b) It is intrinsically a bootstrapping method. Seeding information cannot and need not be supplied, especially not the number of clusters. (c) It has a natural parameter (*inflation*) affecting cluster granularity. (d) It is amenable to sparse graph/matrix implementation techniques, implying good scalability. (e) Mathematical results tie *MCL* process iterands, the cluster interpretation, inflation, and the number of clusters together.

The focus of the present work is largely on (e), the mathematical results describing in a qualitative manner how the *MCL* process exposes cluster structure in graphs. Issues of scaling and implementation are discussed, and in two examples the *MCL* process and its clustering characteristics are visualized. Relationships with other mathematical frameworks are established, and several conjectures are made. Comparison with other clustering approaches fall outside the scope of this exposition. The field of bioinformatics is very active in this respect, and the reader is referred to the references given above.

The *MCL* process is simple to compute and lends itself to drastic scaling by a regime of pruning, as the limits are in general extremely sparse and the iterands sparse in a weighted sense. It is convenient to distinguish between the process and the algorithm, in order to separate mathematical issues from such issues as implementation and scaling (i.e., computing an approximated process in order to gain speed). Section 6 contains a succinct discussion of how an *MCL* implementation can efficiently compute a slightly perturbed *MCL* process.

The structure of the article is as follows. The clustering heuristic is briefly introduced in the next section. The *MCL* process is fully described and the interpretation of a process limit as a clustering of the input graph is given. This is sufficient to define the *MCL* algorithm. A summary is given of some issues concerning convergence and the interpretation of limits as clusterings. Several matrix excerpts from one particular process are shown in section 3, including its limit. In section 4 various lemmas and theorems concerning *MCL* iterands are given. The process consists of alternation of two operators, expansion, and inflation. Both operators preserve the class of stochastic matrices that are diagonally similar to a symmetric matrix. These matrices are called *diagonally symmetric*. Several of their properties are listed. If a matrix is diagonally similar to a positive semi-definite matrix, then it is called a *diagonally positive semi-definite*, abbreviated *dpsd*. Under certain weak conditions many iterands are guaranteed to be *dpsd*. Section 5 introduces structure theory for *dpsd* matrices. Such a matrix possesses structural properties inducing a canonical mapping from the matrix onto a directed acyclic graph, generalizing the mapping from *MCL* limits onto overlapping clusterings. The structure theory also yields a qualitative statement on the working of the inflation operator in terms of the matrix spectrum. Implementation is discussed in section 6, and conclusions, further research, and related research make up the last section.

2. Preliminaries. The *MCL* process consists of alternation of matrix expansion and matrix inflation, where expansion means taking the power of a matrix using the usual matrix product, and inflation (denoted Γ_r) means taking the Hadamard power with coefficient r of a stochastic matrix and subsequently scaling its columns to have sum 1 again. The clustering heuristic associated with the process is that a dense region in a graph corresponds with a node set S for which pairs of elements in S have the property that there are relatively many higher length paths completely contained in S itself. By matrix expansion the higher step transition probabilities are obtained; by matrix inflation large probabilities are promoted, and small probabilities are demoted. It is to be expected that probabilities that correspond with edges connecting different dense regions will suffer the most from the process of alternating expansion and inflation. Indeed, iteration of the two operators leads to a limit that is meaningful considering the original heuristic.

The inflation operator Γ_r is defined for arbitrary nonnegative matrices, in a columnwise manner. This implies that column stochastic matrices will be used rather than row stochastic matrices, which is merely a matter of preference and convention. There are no restrictions on the matrix dimensions to fit a square matrix, because this allows Γ_r to act on both matrices and vectors. There is no restriction that the input matrices be stochastic, since it is not strictly necessary, and the extended applicability is sometimes useful. Following the terminology used in [8] and [27], a nonnegative matrix is called *column allowable* if all its columns have at least one nonzero entry. The next definition prepares for the definition of the *MCL* process.

DEFINITION 2.1. *Denote the operator which raises a square matrix A to the tth power, by* Exp_t. *Thus,* $\mathrm{Exp}_t A = A^t$.

This definition is put in such general terms because the class of dpsd matrices (to be introduced later) allows the introduction of fractional matrix powers in a well-defined way.

DEFINITION 2.2. *Let r be a real positive number, and let $M \in {\mathbb{R}_{\geq 0}}^{m \times n}$ be nonnegative column allowable. The image of M under the parametrized operator Γ_r is defined by setting*

$$(\Gamma_r M)_{pq} \;=\; (M_{pq})^r \,/\, \sum_{i=1}^{m} (M_{iq})^r.$$

In the setting of the *MCL* process, positive values r have a sensible interpretation attached to them. Values of r between 0 and 1 increase the homogeneity of the argument probability vector (matrix), whereas values of r between 1 and ∞ increase the inhomogeneity. In both cases, the ordering of the probabilities is not disturbed. Negative values of r invert the ordering, which is not of apparent use. With $\otimes$ denoting the Kronecker product, the identities $\mathrm{Exp}_r(A \otimes B) = \mathrm{Exp}_r(A) \otimes \mathrm{Exp}_r(B)$ and $\mathrm{Exp}_r(\mathrm{Exp}_s(A)) = \mathrm{Exp}_{rs}(A)$ hold. Similarly, $\Gamma_r(A \otimes B) = \Gamma_r(A) \otimes \Gamma_r(B)$ and $\Gamma_r(\Gamma_s(A)) = \Gamma_{rs}(A)$ are true.

DEFINITION 2.3. *Define Γ_∞ by $\Gamma_\infty M = \lim_{r \to \infty} \Gamma_r M$.*

This definition is meaningful, and it is easy to derive the structure of $\Gamma_\infty M$. Each column q of $\Gamma_\infty M$ has k nonzero entries equal to $1/k$, (k depending on q), where k is the number of elements that equal $\max_p M_{pq}$, and the positions of the nonzero entries in $\Gamma_\infty M[1, \ldots, n|q]$ correspond with the positions of the maximal entries in $M[1, \ldots, n|q]$. Following [44], if x denotes a real vector of length n, then $x_{[1]} \geq x_{[2]} \geq \cdots \geq x_{[n]}$ denote the entries of x in decreasing order.

DEFINITION 2.4. *Let x, y be nonnegative vectors of dimension n. The vector y is said to majorize x, denoted as $x \prec y$, if*

$$\sum_{i=1}^{k} y_{[i]} \geq \sum_{i=1}^{k} x_{[i]} \quad k = 1, \ldots, n, \tag{2.1}$$

$$\sum_{i=1}^{n} y_{[i]} = \sum_{i=1}^{n} x_{[i]}. \tag{2.2}$$

LEMMA 2.5. *For a stochastic vector x and parameters $r, s \in \mathbb{R}_{>0}$, $r < s$, one has that $\Gamma_r(x) \prec \Gamma_s(x)$.*

The proof of this lemma is straightforward [11].

DEFINITION 2.6. *An MCL* process *with input matrix M, where M is a stochastic matrix, is determined by M and two sequences $e_{(i)}$, $r_{(i)}$, where $e_i \in \mathbb{N}, e_i > 1$ and $r_i \in \mathbb{R}, r_i \geq 0$. It is written that*

$$(M, e_{(i)}, r_{(i)}). \tag{2.3}$$

Associated with an MCL process, $(M, e_{(i)}, r_{(i)})$ is an infinite sequence of matrices $M_{(i)}$, where $M_1 = M$, $M_{2i} = \mathrm{Exp}_{e_i}(M_{2i-1})$, and $M_{2i+1} = \Gamma_{r_i}(M_{2i})$, $i = 1, \ldots, \infty$.

It must be stressed that the *MCL* process has no stochastic interpretation. The heuristic on which it is grounded uses stochastic terminology, but each *MCL* process $(M, e_{(i)}, r_{(i)})$ is (for varying M) really a rather complex dynamical system based on the alternation of two operators, expansion and inflation. The fact that expansion and inflation distribute over the Kronecker yields the following lemma.

LEMMA 2.7. *The MCL process distributes over the Kronecker product.*

Note. In practice, clustering with the *MCL* algorithm is best done with all expansion values e_i set to two. The reasoning behind this is pragmatic, as inflation can be used to control the mixing properties of the process, whereas expansion is computationally costly. Applying (columnwise) pruning in order to scale the process renders prolonged expansion virtually useless. Nevertheless it seems best to formulate the *MCL* process in the general terms of Definition 2.6, as this supplies a natural framework for questions and conjectures (section 7). The canonical mapping between graphs with nonnegative weights and nonnegative matrices is given below. In order to work with column stochastic matrices, an arbitrary choice is made to identify matrix columns with lists of neighbors.

DEFINITION 2.8. *The* associated graph *of a square nonnegative matrix A of dimension n is a graph on n nodes labeled $\{1, \ldots, n\}$, where there is said to be an arc going from q to p with weight A_{pq} iff $A_{pq} > 0$.*

The following theorem is preparatory to the mapping from nonnegative idempotent matrices to overlapping clusterings in Definition 2.11. Its proof is given in [11] and can also be derived from the decomposition of nonnegative idempotent matrices given in [2, p. 65]. It represents a very basic result on the structural properties of nonnegative idempotent matrices. Theorem 5.4 will show a more general structure to be present in *MCL* iterands, so that in the setting of the *MCL* process Theorem 2.9 becomes a limiting case of Theorem 5.4. It will be shown that for M stochastic *dpsd* a finite power of the matrix $\Gamma_\infty(M)$ is idempotent (section 5).

THEOREM 2.9 (see Theorem 1 in [11, p. 18]). *Let M be a nonnegative column allowable idempotent matrix of dimension n, and let G be its associated graph. For s, t, nodes in G, write $s \to t$ if there is an arc in G from s to t. By definition,*

$s \to t \iff M_{ts} \neq 0$. Let α, β, γ be nodes in G. The following implications hold:

$$(\alpha \to \beta) \wedge (\beta \to \gamma) \implies \alpha \to \gamma, \tag{2.4}$$

$$(\alpha \to \alpha) \wedge (\alpha \to \beta) \implies \beta \to \alpha, \tag{2.5}$$

$$\alpha \to \beta \implies \beta \to \beta. \tag{2.6}$$

The theorem basically states that the graph associated with the matrix consists for one part of subgraphs that are complete, with all nodes having loops as well. The other part consists of nodes without loops that, given a complete subgraph, are connected either to all or to none of the nodes in that subgraph. It is convenient to introduce the notions of *attractor* and *attractor system*. The second is a (maximal) complete subgraph, and the first is a node in such a subgraph.

DEFINITION 2.10. *Let G be the associated graph of a nonnegative column allowable idempotent matrix M of dimension n, with nodes labeled $1, \ldots, n$. The node α is called an* attractor *if $M_{\alpha\alpha} \neq 0$. If α is an attractor, then the set of nodes reachable from α is called an* attractor system.

By Theorem 2.9, each attractor system in G induces a weighted subgraph in G that is complete. These subgraphs form the cores of the clustering associated with a (nonnegative idempotent) matrix M as stated below. An attractor system is simply extended with all the nodes that reach it.

DEFINITION 2.11. *Let M be a nonnegative column allowable idempotent matrix of dimension n, and let G be its associated graph on the node set $V = \{1, \ldots, n\}$. Let $E_i, i = 1, \ldots, k$ be the different attractor systems of G. For $v \in V$ write $v \to E_i$ if there exists $e \in E_i$ with $v \to e$. The (possibly) overlapping clustering $\mathcal{C} = \{C_1, \ldots, C_k\}$, associated with M, is defined by*

$$C_i \;=\; E_i \cup \{v \in V \mid v \to E_i\}. \tag{2.7}$$

Theorem 2.9 implies that $v \to f$ for all $f \in E_i$.

The simplest example of a limit matrix inducing overlap is the matrix below, giving rise to the clustering $\{1, 3\}, \{2, 3\}$:

$$\begin{pmatrix} 1 & 0 & 1/2 \\ 0 & 1 & 1/2 \\ 0 & 0 & 0 \end{pmatrix}.$$

Combining the previous simple results, it is possible to rewrite each nonnegative column allowable idempotent matrix M as a form $P^T AP$, where P is a permutation matrix, and

$$A = \begin{pmatrix} B_1 & & & f_{11} & f_{12} & \cdots & f_{1l} \\ & \ddots & & \vdots & \vdots & \vdots & \vdots \\ & & B_k & f_{k1} & f_{k2} & \cdots & f_{kl} \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \end{pmatrix}.$$

Each matrix B_i is square, has rank one with all columns identical, contains only positive entries, and there are no other nonzero entries in the corresponding columns of A. Each matrix B_i corresponds with an attractor system, and k is the number of resulting clusters. Each f_{ij} is a column vector with the same number of (row) entries as B_i. Either all entries in f_{ij} are zero or they are all nonzero, and for each j at least one f_{ij} is nonzero. If the vector f_{ij} is nonzero, then it corresponds with a node

(identified by j) that is in the cluster defined by the attractor system corresponding with B_i. If f_{ij} is nonzero for more than one i, then those i determine clusters that overlap in the node identified by j.

In practice, cluster overlap is very rare. The phenomenon is inherently unstable, in the sense that applying the *MCL* process to a perturbation of a limit matrix that induces overlap leads the process to converge to a limit no longer inducing overlap. A node previously in overlap will then be associated with just one of the multiple clusters it was associated with before [12]. All current evidence suggests that cluster overlap implies the existence of a graph automorphism of the graph associated with the input matrix, leaving the overlapping part invariant and mapping the overlapping clusters onto each other. In the simple example above, the automorphism would send $(1, 2, 3)$ to $(2, 1, 3)$.

The phenomenon of attractor systems of cardinality greater than one is also unstable in nature, but a small perturbation of a matrix limit having such a system will not change the associated clustering (assuming that the parameter r of Γ_r is bounded). The main reason for this is that if J is a stochastic matrix of rank one and E is a perturbation matrix (with zero column sums) of sufficiently small norm, then (restricting attention to a special case) $\mathrm{Exp}_2(\Gamma_2(J + E))$ is of the form $J' + E'$, with J' stochastic of rank one and the norm of E' being of order square the norm of E. Current evidence also suggests that attractor systems of cardinality greater than one imply the existence of a set of automorphisms by which each of the attractors (of one system) can be mapped to any of the other. An example is shown in Figure 3.2 for the graph in Figure 3.1. In this case, the automorphism would leave all nodes in place except for interchanging 9 and 11.

Assuming that e_i equals two and r_i is bounded eventually, it is true that the *MCL* process converges quadratically in the neighborhood of matrices that (i) are *MCL-invariant*, that is, invariant both under expansion (multiplication) and inflation, and (ii) have in each column one entry equal to 1 and all other entries equal to 0. This is straightforward (though tedious) to verify—proofs are given in [11]. The issue is somewhat clouded by the fact that the process may also converge towards a limit matrix that does not satisfy condition (ii). A small perturbation of such a matrix is amplified by the inflation operator so that the sequence of iterands departs from it.

The *MCL* algorithm consists of three steps. First, given an arbitrary input graph G, loops are added resulting in a graph informally denoted as $G + \Delta$. Some remarks on the necessity of this step are made in the next section. How weights are chosen for the loops to be added is the responsibility of the algorithm. Subsequently, an *MCL* process is applied to the matrix of random walks associated with $G + \Delta$. Third, the limit thus computed is interpreted as a clustering according to Definition 2.11. One can obtain a fast, robust, and well-scaling implementation of the *MCL* algorithm at http://micans.org/mcl/, which allows a simple type of parametrization: The expansion values e_i are all set to 2 and the inflation values r_i can assume two values, changing once from the first to the second value.

In general the limit of an *MCL* process is extremely sparse, as the inflation operator is a force driving towards sparse columns. *MCL* iterands tend to be sparse in a weighted sense, and this supplies the means to scale the *MCL* algorithm drastically by incorporating a regime of pruning into the *MCL* process (cf. section 6).

The natural way to use the *MCL* process for the purpose of clustering a graph is

by applying it to the matrix which represents the standard concept of a random walk on the graph, where loops have been added to the graph. This matrix is obtained as the incidence matrix multiplied by the diagonal matrix of inverse column (row) sums, so that the product is column (row) stochastic. If the graph is undirected, then the resulting stochastic matrix is diagonally similar to a symmetric matrix.

3. Examples.

Example I. In Figure 3.2, four excerpts are given of an *MCL* process. These are the input matrix M, the iterand $M_3 = \Gamma_2 M^2$, the iterand $M_5 = \Gamma_2(\Gamma_2 M^2 \cdot \Gamma_2 M^2)$, and the stable limit denoted L_M. The process consists entirely of alternation of Exp_2 and Γ_2. The graph H associated with M is depicted in Figure 3.1. Every node in H has a loop; these are all left out in the figure. Weights are omitted as well. Note that there exists a diagonal matrix d such that Md is symmetric. This implies that $d^{-1/2}Md^{1/2}$ is symmetric and thus the spectrum of M is real. Interpreting L_M according to Definition 2.11 yields the clustering $\{\{1,6,7,10\},\{2,3,5\},\{4,8,9,11,12\}\}$. It is necessary to add loops to the nodes before applying *MCL* in order to prevent a result reflecting the bipartite characteristics of H. Without adding loops, the resulting *MCL* process limit yields the clustering $\{\{1,5,10\},\{2,6,7\},\{3,4,8,9,11,12\}\}$. This is in line with the heuristic underlying the process: The probabilities that are initially boosted correspond with 2-step paths in H.

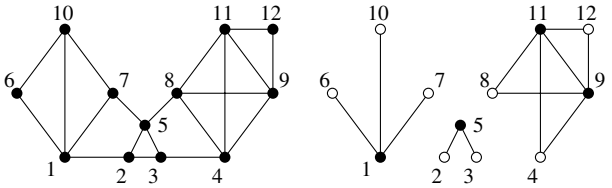

FIG. 3.1. *On the left a graph H, on the right the graph associated with the limit of an MCL process applied to H, loops added to H. Dark circles signify attractors; nodes* 9 *and* 11 *form an attractor system (refer to section* 5*). Compare with the matrix iterands and limit matrix in Figure* 3.2 *and with Figure* 3.3*, and see the discussion in Example* I.

Example II. Figure 3.3 depicts different iterands of an *MCL* process triggered by a geometric graph. This graph was first used in [21] as a test case for graph partitioning. It is shown in the upper left of the figure. Two nodes are connected if their distance is at most $\sqrt{8}$ Euclidean units. The edge weights were taken inversely proportional to the Manhattan distance, and loops were added to each node with a weight equal to the largest weight found in the edges in which it participates. The matrix of random walks on this graph was input to an *MCL* process in which the sequence $e_{(i)}$ assumed the constant 2 everywhere, and the sequence $r_{(i)}$ assumed the constant 1.3 everywhere.

The other graphs in Figure 3.3 represent a pictorial representation of four *MCL* iterands (stochastic matrices) and the limit in the lower right. The degree of shading of a bond between two nodes indicates the maximum value of the corresponding transition probabilities taken over the two directions. The darker the bond, the larger the maximum. The degree of shading of a node indicates the total sum of incoming transition probabilities. Thus, a dark bond between a white node and a black node indicates that the maximum transition probability is found in the direction of the black node, and that the probability attached to the reverse arc is negligible. The limit graph, depicted in the lower right, contains all necessary information needed for

$$\begin{pmatrix}
0.200 & 0.250 & -- & -- & -- & 0.333 & 0.250 & -- & -- & 0.250 & -- & -- \\
0.200 & 0.250 & 0.250 & -- & 0.200 & -- & -- & -- & -- & -- & -- & -- \\
-- & 0.250 & 0.250 & 0.200 & 0.200 & -- & -- & -- & -- & -- & -- & -- \\
-- & -- & 0.250 & 0.200 & -- & -- & -- & 0.200 & 0.200 & -- & 0.200 & -- \\
-- & 0.250 & 0.250 & -- & 0.200 & -- & 0.250 & 0.200 & -- & -- & -- & -- \\
0.200 & -- & -- & -- & -- & 0.333 & -- & -- & -- & 0.250 & -- & -- \\
0.200 & -- & -- & -- & 0.200 & -- & 0.250 & -- & -- & 0.250 & -- & -- \\
-- & -- & -- & 0.200 & 0.200 & -- & -- & 0.200 & 0.200 & -- & 0.200 & -- \\
-- & -- & -- & 0.200 & -- & -- & -- & 0.200 & 0.200 & -- & 0.200 & 0.333 \\
0.200 & -- & -- & -- & -- & 0.333 & 0.250 & -- & -- & 0.250 & -- & -- \\
-- & -- & -- & 0.200 & -- & -- & -- & 0.200 & 0.200 & -- & 0.200 & 0.333 \\
-- & -- & -- & -- & -- & -- & -- & -- & 0.200 & -- & 0.200 & 0.333
\end{pmatrix}$$

M

$$\begin{pmatrix}
0.380 & 0.087 & 0.027 & -- & 0.077 & 0.295 & 0.201 & -- & -- & 0.320 & -- & -- \\
0.047 & 0.347 & 0.210 & 0.017 & 0.150 & 0.019 & 0.066 & 0.011 & -- & 0.012 & -- & -- \\
0.014 & 0.210 & 0.347 & 0.055 & 0.150 & -- & 0.016 & 0.046 & 0.009 & -- & 0.009 & -- \\
-- & 0.027 & 0.087 & 0.302 & 0.062 & -- & -- & 0.184 & 0.143 & -- & 0.143 & 0.083 \\
0.058 & 0.210 & 0.210 & 0.055 & 0.406 & -- & 0.083 & 0.046 & 0.009 & 0.019 & 0.009 & -- \\
0.142 & 0.017 & -- & -- & -- & 0.295 & 0.083 & -- & -- & 0.184 & -- & -- \\
0.113 & 0.069 & 0.017 & -- & 0.062 & 0.097 & 0.333 & 0.011 & -- & 0.147 & -- & -- \\
-- & 0.017 & 0.069 & 0.175 & 0.049 & -- & 0.016 & 0.287 & 0.143 & -- & 0.143 & 0.083 \\
-- & -- & 0.017 & 0.175 & 0.012 & -- & -- & 0.184 & 0.288 & -- & 0.288 & 0.278 \\
0.246 & 0.017 & -- & -- & 0.019 & 0.295 & 0.201 & -- & -- & 0.320 & -- & -- \\
-- & -- & 0.017 & 0.175 & 0.012 & -- & -- & 0.184 & 0.288 & -- & 0.288 & 0.278 \\
-- & -- & -- & 0.044 & -- & -- & -- & 0.046 & 0.120 & -- & 0.120 & 0.278
\end{pmatrix}$$

$\Gamma_2 M^2$

$$\begin{pmatrix}
0.448 & 0.080 & 0.023 & 0.000 & 0.068 & 0.426 & 0.359 & 0.000 & 0.000 & 0.432 & 0.000 & -- \\
0.018 & 0.285 & 0.228 & 0.007 & 0.176 & 0.006 & 0.033 & 0.005 & 0.000 & 0.007 & 0.000 & 0.000 \\
0.005 & 0.223 & 0.290 & 0.022 & 0.173 & 0.000 & 0.010 & 0.017 & 0.003 & 0.001 & 0.003 & 0.001 \\
0.000 & 0.018 & 0.059 & 0.222 & 0.040 & 0.000 & 0.001 & 0.187 & 0.139 & 0.000 & 0.139 & 0.099 \\
0.027 & 0.312 & 0.314 & 0.028 & 0.439 & 0.005 & 0.054 & 0.022 & 0.003 & 0.010 & 0.003 & 0.001 \\
0.116 & 0.007 & 0.001 & 0.000 & 0.004 & 0.157 & 0.085 & 0.000 & -- & 0.131 & -- & -- \\
0.096 & 0.040 & 0.013 & 0.000 & 0.037 & 0.083 & 0.197 & 0.001 & 0.000 & 0.104 & 0.000 & 0.000 \\
0.000 & 0.012 & 0.042 & 0.172 & 0.029 & 0.000 & 0.002 & 0.198 & 0.133 & 0.000 & 0.133 & 0.096 \\
0.000 & 0.001 & 0.015 & 0.256 & 0.009 & -- & 0.000 & 0.266 & 0.326 & 0.000 & 0.326 & 0.346 \\
0.290 & 0.021 & 0.002 & 0.000 & 0.017 & 0.323 & 0.260 & 0.000 & 0.000 & 0.316 & 0.000 & -- \\
0.000 & 0.001 & 0.015 & 0.256 & 0.009 & -- & 0.000 & 0.266 & 0.326 & 0.000 & 0.326 & 0.346 \\
-- & 0.000 & 0.001 & 0.037 & 0.000 & -- & 0.000 & 0.039 & 0.069 & -- & 0.069 & 0.112
\end{pmatrix}$$

$\Gamma_2(\Gamma_2 M^2 \cdot \Gamma_2 M^2)$

$$\begin{pmatrix}
1.000 & -- & -- & -- & -- & 1.000 & 1.000 & -- & -- & 1.000 & -- & -- \\
-- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- \\
-- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- \\
-- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- \\
-- & 1.000 & 1.000 & -- & 1.000 & -- & -- & -- & -- & -- & -- & -- \\
-- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- \\
-- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- \\
-- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- \\
-- & -- & -- & 0.500 & -- & -- & -- & 0.500 & 0.500 & -- & 0.500 & 0.500 \\
-- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- \\
-- & -- & -- & 0.500 & -- & -- & -- & 0.500 & 0.500 & -- & 0.500 & 0.500 \\
-- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- & --
\end{pmatrix}$$

Limit L_M resulting from iterating $(\Gamma_2 \circ \mathrm{Exp}_2)$ with initial matrix M, which is the matrix of random walks associated with the graph in Figure 3.1.

Entries marked "−−" are either zero because that is the exact value they assume (this is true for the first two matrices) or because the computed value fell below the machine precision.

FIG. 3.2. *Iteration of* $(\Gamma_2 \circ \mathrm{Exp}_2)$.

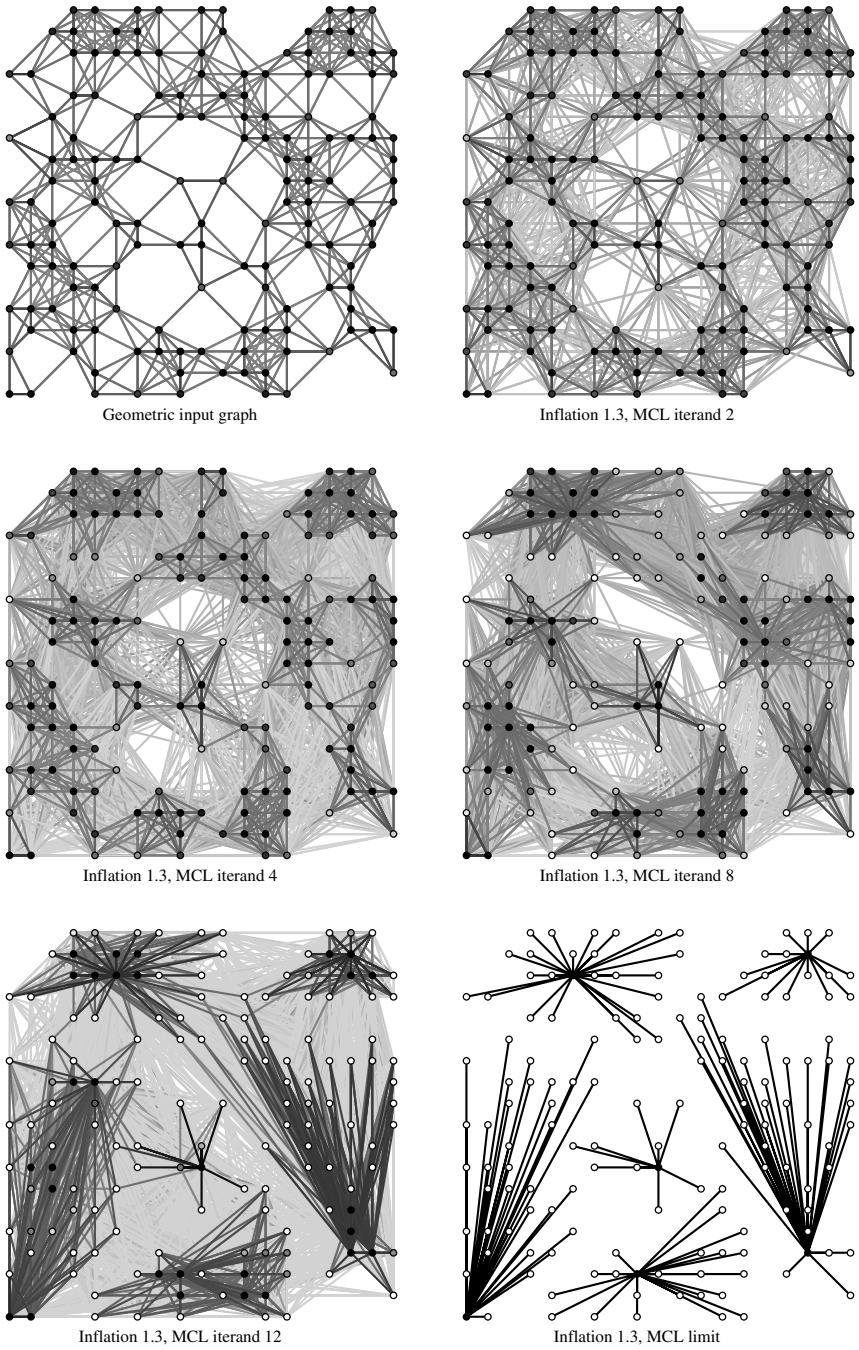

FIG. 3.3. *Visualization of successive stages of the MCL process applied to the upper left graph, with* $e_i = 2$ *and* $r_i = 1.3$ *for every iteration* i *(cf. Definition* 2.6*). The meaning of the grey values of bonds and nodes are explained in section* 3*. At most* 24 *neighbors are shown for each node.*

constructing the *MCL*-invariant limit matrix. Dark nodes in this graph are attractors.

The examples in Figures 3.3 and 3.2 indicate that the *MCL* process has remarkable convergence properties, regarding the structural properties of its iterands. Considering this evidence, to some extent an analogy is suggested with the normal Markov process.

Assuming that the associated graph of the input matrix M is strongly connected and contains at least one loop, it follows by Perron–Frobenius theory that 1 is the only eigenvalue of M of modulus 1 and that it is simple.

By considering the spectrum of the powers M^k it follows that the normal Markov process converges towards a rank-one idempotent matrix, having spectrum $\{0^{n-1}, 1\}$. In the example shown in Figure 3.2 the process also converges towards an idempotent limit. The multiplicity of its eigenvalue 1 is 3, however, equaling (of course) the number of strongly connected components in the associated graph of the limit. Section 4 will give some insight into the spectral phenomena that play a role in the *MCL* process by focusing attention onto two classes of stochastic matrices.

4. Properties of the inflation operator and stochastic *dpsd* matrices. At first sight the inflation operator seems hard to get a grasp on mathematically, though its behavior for vectors is well understood. Lemma 2.5 states that for a stochastic vector x and parameters $r, s \in \mathbb{R}_{>0}$, $r < s$, one has that $\Gamma_r(x) \prec \Gamma_s(x)$, where $\prec$ denotes the majorization relationship. This implies that the orbit $\Gamma_r x$, $(r > 0)$ is fairly well understood, since the limiting cases $\Gamma_r x$, $r \to \infty$ and $\Gamma_r x$, $r \downarrow 0$ are also easily derived. However, majorization results for vectors do not carry over to matrices in such a way that statements can be made about algebraic properties of two matrices subject to a columnwise majorization relationship. In [44] this issue is discussed at length.

To some extent it is possible to give a qualitative account of the behavior of the inflation operator, using structural properties of the matrices in a particular class preserved by inflation. Several preparatory results are derived in the current section. In the following section simple structure theory is developed, explaining the uncoupling effect of the inflation operator in qualitative terms.

In general $\Gamma_r M$ can be described in terms of a Hadamard matrix power that is postmultiplied with a diagonal matrix. For a restricted class of matrices there is an even stronger connection with the Hadamard product. These are the class of stochastic diagonally symmetric matrices and a subclass of the latter, the class of stochastic diagonally positive semi-definite matrices.

The Hadamard (entrywise) product of two matrices A and B that have the same dimensions is written $A \circ B$ and satisfies $[A \circ B]_{pq} = A_{pq} B_{pq}$. The entrywise Hadamard power with exponent r of a matrix A is written $A^{\circ r}$ and satisfies $[A^{\circ r}]_{pq} = {A_{pq}}^r$.

The concept of diagonal symmetrizability can easily be transferred to complex matrices, and most of the results in this paper can be derived in that more general setting. This is not needed in the *MCL* setting and hence the definitions and results here are simply stated for real matrices.

DEFINITION 4.1. *A square matrix A is called* diagonally symmetric *if it is diagonally similar to a symmetric matrix, that is, if there exists a positive vector x such that the product* $\mathrm{Diag}(x)^{-1} A \,\mathrm{Diag}(x)$ *is symmetric.*

The following useful identity is easy to verify.

LEMMA 4.2. *For a matrix A as in Definition* 4.1*, the identity* $\mathrm{Diag}(x)^{-1} A \,\mathrm{Diag}(x) = [A \circ A^T]^{\circ 1/2}$ *holds.*

DEFINITION 4.3. *A square matrix is called* diagonally positive semi-definite *if it is diagonally similar to a positive semi-definite matrix, then it is called* diagonally positive definite *if it is diagonally similar to a positive definite matrix. The phrases are respectively abbreviated as* dpsd *and* dpd.

Remark. If M is diagonally symmetric stochastic, and y is such that $M\ \mathrm{Diag}(y)$ is symmetric, then $My = y$; thus y represents the equilibrium distribution of M. In the theory of Markov chains, a stochastic diagonally symmetric matrix is called *time reversible* or said to satisfy the *detailed balance* condition (see, e.g., [43] and [59]). A slightly more general definition and different terminology was chosen here. The main reason is that the term "time reversible" is coupled tightly with the idea of studying a stochastic chain via (powers of) its associated stochastic matrix, and is also used for continuous-time Markov chains. The *MCL* process studied in this article does not have a straightforward stochastic interpretation, and the relationship between an input matrix and the subsequent iterands is much more complex. Moreover, it is natural to introduce the concepts of a matrix being diagonally similar to a positive (semi-) definite matrix; clinging to "time reversible" in this abstract setting would be both contrived and unhelpful. The proposed phrases seem appropriate, since several properties of symmetric and *psd* matrices remain valid in the more general setting of diagonally symmetric and *dpsd* matrices. Lemma 4.4 lists the most important ones, which are easy to verify. Probably all of these results are known.

In the following, submatrices of a matrix A are written $A[u|v]$, where u denotes a list of row indices, and v denotes a list of column indices.

LEMMA 4.4. *Let A be diagonally symmetric of dimension n, let α be a list of distinct indices in the range $1, \ldots, n$, and let k and l be different indices in the range $1, \ldots, n$. Let x be such that $S = \mathrm{Diag}(x)^{-1} A\ \mathrm{Diag}(x)$ is symmetric, and thus $A = \mathrm{Diag}(x) S\ \mathrm{Diag}(x)^{-1}$. Let λ_i be the eigenvalues of A (and S), and let a_i be the diagonal entries of A.*

(a) *$A[\alpha|\alpha] = \mathrm{Diag}(x)[\alpha|\alpha]\ \ S[\alpha|\alpha]\ \ \mathrm{Diag}(x)[\alpha|\alpha]^{-1}$, in particular, the diagonal entries of A equal the diagonal entries of S. This implies that the majorization relationship between eigenvalues and diagonal entries for symmetric matrices carry over to diagonally symmetric matrices: The spectrum of A majorizes the vector of diagonal entries of A, translating to the inequalities below:*

$$\sum_{i=1}^{k} \lambda_{[i]} \ \geq \ \sum_{i=1}^{k} a_{[i]} \quad k = 1, \ldots, n.$$

Together with the first equality this implies that diagonally symmetric matrices satisfy the same interlacing inequalities for bordered matrices as symmetric matrices do.

(b) *If A is* dpsd *and $A_{kk} = 0$, then the kth row and the kth column of A are zero. If A is* dpsd *and $\det A[kl|kl] = 0$, then row k and row l are proportional, and column k and column l are proportional.*

(c) *If A is* dpsd*, then, for each $k \in \mathbb{N}$, there exists a unique* dpsd *matrix B such that $B^k = A$. This matrix is defined by setting $B = \mathrm{Diag}(x) Q \Lambda^{1/k} Q^H\ \mathrm{Diag}(x)^{-1}$, where $Q \Lambda Q^H$ is a unitary diagonalization of S, Λ is the diagonal matrix of eigenvalues of S, and $\Lambda^{1/k}$ is the matrix Λ with each diagonal entry replaced by its real nonnegative kth root. This implies that for* dpsd *A, the fractional power A^t, $t \in \mathbb{R}_{\geq 0}$, can be defined in a meaningful way.*

(d) *If A, B are both of dimension n and diagonally symmetric,* dpsd, dpd*, then the Hadamard product $A \circ B$ is diagonally symmetric,* dpsd, dpd.

Proof. Most statements are easy to verify. For extensive discussion of the majorization relationship between diagonal entries and eigenvalues of symmetric (or hermitian) matrices, as well as results on interlacing inequalities, see [3, 34, 35]. The first statement in (b) follows from the fact that principal minors (of dimension 2) are nonnegative. The second statement can easily be proven by first considering the case where A is symmetric. The determinant $\det A[klm|klm]$ of an extended submatrix equals zero and rewriting the constituent terms yields proportionality as stated in (b). The result for *dpsd* matrices follows trivially. For (c) it is sufficient to use the fact that $Q\Lambda^{1/k}Q^H$ is the unique positive semi-definite kth root of S [34, p. 405]. Statement (d) follows from the identity $(\mathrm{Diag}(x)^{-1}A\ \mathrm{Diag}(x)) \circ (\mathrm{Diag}(y)^{-1}B\ \mathrm{Diag}(y)) = \mathrm{Diag}(x \circ y)^{-1}(A \circ B)\mathrm{Diag}(x \circ y)$ and the fact that the analogous statements for symmetric matrices are true—known under the denomination of *Schur product theorem* [34, p. 458].

Remark. The two most notable properties that do not generalize from symmetric matrices to diagonally symmetric matrices are the absence of an orthogonal basis of eigenvectors for the latter, and the fact that the sum of two diagonally symmetric matrices is in general not diagonally symmetric as well.

Statements (a) and (b) in Lemma 4.4 are used in associating a directed acyclic graph with each *dpsd* matrix in Theorem 5.4. First, the behavior of the inflation operator on diagonally symmetric and *dpsd* matrices is described.

THEOREM 4.5. *Let M be a square column allowable diagonally symmetric matrix of dimension n, and let* $\mathrm{Diag}(x)$ *be the diagonal matrix with a positive diagonal such that the matrix $S = \mathrm{Diag}(x)^{-1}M\ \mathrm{Diag}(x)$ is symmetric, and let r be real. Define the positive vector z by setting $z_k = {x_k}^r(\sum_i {M_{ik}}^r)^{1/2}$, and the positive rank-one symmetric matrix T by setting $T_{kl} = 1/(\sum_i {M_{ik}}^r)^{1/2}(\sum_i {M_{il}}^r)^{1/2}$. The following statement holds:*

$$\mathrm{Diag}(z)^{-1}\ (\ \Gamma_r M)\ \ \mathrm{Diag}(z)\ \ =\ \ S^{\circ r} \circ T.$$

Thus $\Gamma_r M$ is diagonally similar to a symmetric matrix.

Proof. Define the vector t by $t_k = \sum_i {M_{ik}}^r$. Then

$$\begin{aligned}
\Gamma_r M =\ & M^{\circ r}\ \mathrm{Diag}(t)^{-1} \\
=\ & (\mathrm{Diag}(x)\ S\ \mathrm{Diag}(x)^{-1})^{\circ r}\ \mathrm{Diag}(t)^{-1} \\
=\ & \mathrm{Diag}(x)^{\circ r}\ \ S^{\circ r}\ \ (\mathrm{Diag}(x)^{\circ r})^{-1}\ \mathrm{Diag}(t)^{-1} \\
=\ & \mathrm{Diag}(t)^{1/2}\mathrm{Diag}(t)^{-1/2}\ \mathrm{Diag}(x)^{\circ r}\ \ S^{\circ r}\ (\mathrm{Diag}(x)^{\circ r})^{-1}\mathrm{Diag}(t)^{-1/2}\mathrm{Diag}(t)^{-1/2} \\
=\ & (\mathrm{Diag}(t)^{1/2}\mathrm{Diag}(x)^{\circ r})\ (\mathrm{Diag}(t)^{-1/2}S^{\circ r}\mathrm{Diag}(t)^{-1/2})\ (\mathrm{Diag}(t)^{1/2}\mathrm{Diag}(x)^{\circ r})^{-1}.
\end{aligned}$$

Since the matrix $\mathrm{Diag}(t)^{-1/2}S^{\circ r}\mathrm{Diag}(t)^{-1/2}$ equals $S^{\circ r} \circ T$, the lemma holds. □

COROLLARY 4.6. *Let M be square column allowable diagonally symmetric, and let z, S, and T be as in Theorem* 4.5.

(i) *The matrix $\Gamma_r M$ is diagonally symmetric for all $r \in \mathbb{R}$.*

(ii) *If M is* dpsd*, then $\Gamma_r M$ is* dpsd *for all $r \in \mathbb{N}$. If M is* dpd*, then $\Gamma_r M$ is* dpd *for all $r \in \mathbb{N}$.*

Proof. Statement (i) follows immediately from Theorem 4.5. Statement (ii) follows from the fact that a Hadamard product of matrices is positive (semi-) definite if

each of the factors is positive (semi-) definite. Moreover, if at least one of the factors is positive definite, and none of the other factors has a zero diagonal entry, then the product is positive definite (see, e.g., [35, p. 309], or [23]). These are basic results in the theory of Hadamard products, an area now covered by a vast body of literature. Standard references in this area are [3, 35]. It should be noted that $r \in \mathbb{N}$ is in general a necessary condition [35, p. 453]. □

THEOREM 4.7. *Let M be diagonally symmetric stochastic, and consider the MCL process $(M, e_{(i)}, r_{(i)})$.*

(i) *All iterands of this process have real spectrum.*

(ii) *If $r_i = 2$ eventually, and $e_i = 2$ eventually, then the iterands of the process $(M, e_{(i)}, r_{(i)})$ are* dpsd *eventually.*

These statements[1] follow from the fact that Exp_2 maps diagonally symmetric matrices onto *dpsd* matrices and from Corollary 4.6. □

Theorem 4.7 represents a qualitative result on the *MCL* process. Under fairly basic assumptions the spectra of the iterands are real and nonnegative. In [11] it was furthermore proven that the *MCL* process converges quadratically in the neighborhood of nonnegative *MCL*-invariant matrices. These combined facts indicate that the *MCL* process has a sound mathematical foundation. Still, much less can be said about the connection between successive iterands than in the case of the discrete Markov process.

The question now rises whether the *MCL* process can be further studied aiming at quantitative results. It was seen that $\Gamma_r M$, $r \in \mathbb{N}$ can be described in terms of a Hadamard product of positive semi-definite matrices relating the symmetric matrices associated with M and $\Gamma_r M$ (in Theorem 4.5). There are many results on the spectra of such products. The results are generically in terms of a majorization relationship such as

$$\sum_{i=1}^{k} \sigma_i(A \circ B) \;\leq\; \sum_{i=1}^{k} f_i(A)\sigma_i(B), \quad k = 1, \ldots, n.$$

Here $\sigma_i()$ denotes the i-largest singular value, and $f_i(A)$ may stand (among others) for the i-largest singular value of A, the i-largest diagonal entry of A, the i-largest Euclidean column length, or the i-largest Euclidean row length. Well-known references in this field are [3, 35]. Unfortunately such inequalities go the wrong way in a sense. Since the inflation operator has apparently the ability to press several large eigenvalues towards 1, what is needed are inequalities of the type

$$\sum_{i=1}^{k} \sigma_i(A \circ B) \;\geq\; \text{(something nice here)}.$$

However, the number of eigenvalues pressed towards 1 by Γ_r can be any number including zero (noting that one eigenvalue 1 is always present). Moreover, Γ_r also has the ability to press small eigenvalues towards zero. Clearly, one cannot expect to find inequalities of the "$\geq$" type without assuming additional characteristics of M. It is shown in the next section that the classic majorization relation formulated in

[1] Clearly the condition under (ii) can be weakened; it is only necessary that e_i is at least one time even for an index $i = k$ such that $r_i \in \mathbb{N}$ for $i \geq k$. However, the assumptions under (ii) can be viewed as a standard way of enforcing convergence in a setting genuinely differing from the discrete Markov process.

Lemma 4.4 (a) between the eigenvalues and diagonal entries of a *dpsd* matrix, plus a classification of the diagonal entries of a *dpsd* matrix, gives useful information on the relationship between eigenvalues of a stochastic *dpsd* matrix and its image under Γ_r.

5. Structure in *dpsd* matrices. The main objective for the remainder of this paper is to establish structure theory for the class of *dpsd* matrices and study the behavior of Γ_∞ using these results. It will be shown that for stochastic *dpsd* M the spectrum of the matrix Γ_∞ is of the form $\{0^{n-k}, 1^k\}$, where k is related to a structural property of M. Throughout this section two symbols are used that are associated with a *dpsd* matrix A, namely the symbol $\not\mapsto$ which denotes an arc relation defined on the indices of A, and the symbol $\sim$ which denotes an equivalence relation on the indices of A. It should be clear from the context which matrix they refer to. All results in this section are stated in terms of columns; the analogous statements in terms of rows hold as well.

DEFINITION 5.1. *Let A be* dpsd *of dimension n, and let k and l be different indices in the range $1, \ldots, n$.*

(i) *Define the equivalence relation $\sim$ on the set of indices $\{1, \ldots, n\}$ by $k \sim l \equiv$* columns k and l of A are scalar multiples of each other via scalars on the complex unit circle.

(ii) *Define the arc relation $\not\mapsto$ on the set of indices $\{1, \ldots, n\}$, for $p \neq q$, by $q \not\mapsto p \equiv |A_{pq}| \geq |A_{qq}|$.*

(iii) *Let E and F be different equivalence classes in $\{1, \ldots, n\}/\sim$. Extend the definition of $\not\mapsto$ by setting $F \not\mapsto E \equiv \exists e \in E, \exists f \in F[f \not\mapsto e]$. By definition of $\not\mapsto$ and $\sim$, the latter implies that $\forall e' \in E, \forall f' \in F[f' \not\mapsto e']$.*

LEMMA 5.2. *Let A be* dpsd *of dimension n, and let k and l be distinct indices in the range $1, \ldots, n$. Then*

$$l \not\mapsto k \wedge k \not\mapsto l \text{ implies } k \sim l.$$

This follows from Lemma 4.4 (b) and the fact that the assumption implies $\det A[kl|kl] = 0$. The following lemma prepares for a mapping of *dpsd* matrices onto directed acyclic graphs.

LEMMA 5.3. *Let A be* dpsd *of dimension n, suppose there exist k distinct indices $p_i, i = 1, \ldots, k, k > 1$, such that $p_1 \not\mapsto p_2 \not\mapsto \ldots \not\mapsto p_k \not\mapsto p_1$. Then $p_1 \sim p_2 \sim \ldots \sim p_k$, and thus all $p_i, i = 1, \ldots, k$ are contained in the same equivalence class in $\{1, \ldots, n\}/\sim$. Furthermore, if A is real nonnegative, then the subcolumns $A[p_1 \ldots p_k|p_i]$ are a scalar multiple of the all-one vector of length k.*

Proof. Without loss of generality, assume $1 \not\mapsto 2 \not\mapsto \ldots \not\mapsto k \not\mapsto 1$. The following inequalities hold, where the left-hand side inequalities follow from the inequalities implied by $\det A[i\, i{+}1] \geq 0$ and $i \not\mapsto i+1$,

$$\begin{array}{ccccc} |A_{i\,i+1}| & \leq & |A_{i+1\,i+1}| & \leq & |A_{i+2\,i+1}| \\ |A_{k-1\,k}| & \leq & |A_{kk}| & \leq & |A_{1k}| \\ |A_{k1}| & \leq & |A_{11}| & \leq & |A_{21}|. \end{array}$$

Now let x be positive such that $x_q A_{pq} = x_p A_{qp}$. On the one hand, $|A_{kk}| \leq |A_{1k}|$. On

the other hand,

$$\begin{aligned} |A_{kk}| &\geq |A_{k-1\,k}| \\ &= \tfrac{x_{k-1}}{x_k}|A_{k\,k-1}| \\ &\geq \tfrac{x_{k-1}}{x_k}|A_{k-2\,k-1}| \\ &= \tfrac{x_{k-1}}{x_k}\tfrac{x_{k-2}}{x_{k-1}}|A_{k-1\,k-2}| \\ &\cdots \\ &\geq \tfrac{x_{k-1}}{x_k}\tfrac{x_{k-2}}{x_{k-1}}\cdots\tfrac{x_1}{x_2}|A_{k1}| \\ &= \tfrac{x_1}{x_k}|A_{k1}| \\ &= |A_{1k}|. \end{aligned}$$

This implies that $|A_{k-1\,k}| = |A_{kk}| = |A_{1k}|$ and the identities $|A_{i-1\,i}| = |A_{ii}| = |A_{i+1\,i}|$ are established by abstracting from the index k. From this it follows that $\det A[i, i+1|i, i+1] = 0$, and consequently $i \sim i+1$ for $i = 1, \ldots, k-1$ by Lemma 5.2. The identities $|A_{i-1\,i}| = |A_{ii}| = |A_{i+1\,i}|$ also imply the last statement of the lemma. ∎

Lemma 5.2 can now be generalized towards Theorem 5.4.

THEOREM 5.4. *Let A be* dpsd *of dimension n.*

The arc $\looparrowright$ defines a directed acyclic graph (DAG) on $\{1, \ldots, n\}/\sim$.

Note that the theorem is stated in a columnwise manner. The analogous statement for rows is of course also true. The proof of this theorem follows from Lemma 5.3.

THEOREM 5.5. *Let M be stochastic* dpsd *of dimension n. Let D be the directed graph associated with $\Gamma_\infty M$ defined on $\{1, \ldots, n\}/\sim$ according to Definition* 5.1*, which is acyclic according to Theorem* 5.4*. Let d be the depth of D, that is, the length of a longest path in D. Let k be the number of nodes in $\{1, \ldots, n\}/\sim$ which do not have an outgoing arc in D. These nodes correspond with (groups of) indices p for which M_{pp} is maximal in column p.*

The spectrum of $\Gamma_\infty M$ equals $\{0^{n-k}, 1^k\}$.
The matrix $(\Gamma_\infty M)^d$ is idempotent.

Proof. For the duration of this proof, write S_A for the symmetric matrix to which a diagonally symmetric matrix A is similar. For the first statement, consider the identity

$$S_{(\Gamma_r M)} = \left[\Gamma_r M \circ (\Gamma_r M)^T\right]^{\circ 1/2}$$

given in Lemma 4.2. The matrices $\Gamma_r M$ and $S_{\Gamma_r M}$ have the same spectrum. Now, let r approach infinity. The identity is in the limit not meaningful, since $\Gamma_\infty M$ is not necessarily diagonalizable, and thus the left-hand side may not exist in the sense that there is no symmetric matrix to which $\Gamma_\infty M$ is similar. However, the identity [spectrum of $\Gamma_\infty M$ = spectrum of $\left[\Gamma_\infty M \circ (\Gamma_\infty M)^T\right]^{\circ 1/2}$] *does* remain true, since the spectrum depends continuously on matrix entries [34, p. 540], and both limits exist. Thus, it is sufficient to compute the spectrum of S_∞, which is defined as

$$S_\infty = \left[\Gamma_\infty M \circ (\Gamma_\infty M)^T\right]^{\circ 1/2}.$$

Note that the nonzero entries of $\Gamma_\infty M$ correspond with the entries of M which are maximal in their column. Whenever $[\Gamma_\infty M]_{kl} \neq 0$ *and* $[\Gamma_\infty M]_{lk} \neq 0$, it is true that $k \looparrowright l$ and $l \looparrowright k$. Now consider a column q in S_∞, and assume that all nonzero entries in column q of S_∞ are enumerated $S_{\infty p_i q} \neq 0$, for $i = 1, \ldots, t$. It follows that $q \looparrowright p_i \wedge p_i \looparrowright q$ for all i, thus $q \sim p_i$ for all i, and $S_\infty[p_1 \ldots p_t | p_1 \ldots p_t]$ is a positive submatrix equal to $t^{-1}J_t$, where J_t denotes the all-one matrix of dimension t. This implies that S_∞ is block diagonal (after permutation), with each block corresponding with an equivalence class in $\{1, \ldots, n\}/\sim$ which has no outgoing arc in the $\looparrowright$ arc relation. Each block contributes an eigenvalue 1 to the spectrum of S_∞. Since the spectrum of S_∞ equals the spectrum of $\Gamma_\infty M$, and there are assumed to be k equivalence classes with the stated properties, this proves the first statement.

A second approach proves both the first and the second statement. Consider $\Gamma_\infty M$ and the DAG D associated with it. Each index i for which $[\Gamma_\infty M]_{ii} \neq 0$ must be in an endclass of D because Γ_∞ annihilates all but the maximal elements in each column. Moreover, the nonzero diagonal block (possibly 1-dimensional) associated with such an index is idempotent. This implies that $\Gamma_\infty M$ can be decomposed into an idempotent part (consisting of the diagonal block) and a nilpotent part (the rest). Some calculations now verify that $(\Gamma_\infty M)^d$ is idempotent, where d is the depth of D. ∎

Theorems 5.4 and 5.5 shed light on the structure and the spectral properties of the iterands of the *MCL* process. Theorem 5.4 also gives the means to associate an overlapping clustering with each *dpsd* iterand of an *MCL* process, simply by defining the endnodes of the associated *DAG* as the unique cores of the clustering, and adding to each set of attractors all nodes that reach it.

Consider a discrete Markov process with *dpsd* input matrix M. Then the difference $M^k - M^l$, $k < l$, is again *dpsd* (they have the same symmetrizing diagonal matrix, and the spectrum of $M^k - M^l$ is nonnegative). From this it follows that all sequences of diagonal entries ${M^{(k)}}_{ii}$, for fixed diagonal position ii, are nonincreasing. In contrast, given a stochastic *dpsd* matrix M, the Γ_r operator, $r > 1$, (in the setting of *dpsd* matrices) always increases some diagonal entries (at least one). The sum of the increased diagonal entries, of which there are at least k if k is the number of endnodes of the *DAG* associated with both M and $\Gamma_r M$, is a lower bound for the sum of the k largest eigenvalues of $\Gamma_r M$ (see Lemma 4.4 (a)).

The *MCL* process converges quadratically in the neighborhood of the *MCL*-invariant stable states. Proving (near-) global convergence seems to be a difficult task. I do believe, however, that a strong result will hold, where a provision has to be made for a special class of matrices, here dubbed flip-flop matrices. A flip-flop matrix M satisfies $\Gamma_2 M = M^{1/2}$. There exists a family of positive semi-definite flip-flop matrices of the form $aI_n + (1-a)n^{-1}J_n$, $n \in \mathbb{N}$ [12]. The simplest example is found in the case $n = 3$, where substituting $a = 1/2$ in the form yields a flip-flop matrix. For such a matrix it is relatively easy to prove that a small perturbation lands it on a trajectory away from the flip-flop state (with respect to alternation of Exp_2 and Γ_2) [12]. It can be noted that flip-flop matrices and circulant matrices in general form sets that are invariant under *MCL* iterations.

CONJECTURE 1. *All MCL processes* $(M, e_{(i)}, r_{(i)})$*, with* $e_i = 2, r_i = 2$ *eventually, converge towards an MCL-invariant limit, provided* M *is irreducible, stochastic,* dpsd*, and cannot be decomposed as a Kronecker product of matrices in which one of the terms is a flip-flop matrix.*

The requirement of irreducibility is present in order to exclude matrices that are

a direct sum of smaller-dimensional matrices.

6. Implementation and scalability. A mature C implementation of the MCL algorithm is available from http://micans.org/mcl/. This implementation is used in all of the references cited in the introduction. It scales subquadratically given conditions set forth below.

A fast implementation requires that the requirement of exact computation is dropped. For any interesting class of real-life graphs scaling towards tens of thousands of nodes and beyond, exact computation requires $O(N^2)$ memory resources and $O(N^3)$ time steps, where N is the number of nodes in the input graph, reflecting the basic costs of matrix multiplication. Even for sparse graphs, *MCL* iterands will fill rapidly as interesting graphs tend to be well-connected and have only few connected components.

The key observation is that in the presence of cluster structure, columns of *MCL* iterands generally possess a very skewed (weight) distribution of entries. The majority of the stochastic mass of any column is contained in a minority of the total set of nonzero entries (of that column), as inflation keeps the leveling power of expansion (multiplication) in check. In the *MCL* process limits, the matrix columns generally are extremely skewed, with a single nonzero entry per column (equaling one). This implies that *MCL* iterands never stray very far from the skewed weight distribution just described, and it suggests a way to compute a perturbed process that is tractable. That is to simply throw away some of the smallest entries, preferably adding to only a small percentage of the column weight, and rescale the remaining entries to have sum one again. This is the setup in the implementation described here.

The implementation uses a standard sparse matrix implementation where only nonzero entries are stored in arrays representing stochastic matrix columns (known as compressed column or column-major storage). During matrix multiplication, each new column is computed separately. First, the new column is computed exactly and nothing is disregarded. Then, the smallest entries are removed in a two-stage process where first entries smaller than a fixed threshold are removed, and then entries are recovered if the threshold turns out to be too severe, or more entries are removed if the threshold turns out to be insufficiently severe. The selection and recovery of entries is efficiently done using max and min heaps. The final assembly of entries is rescaled to have sum one. The implementation tracks how much mass is kept for each column during each iteration, and extensively reports on pruning characteristics.

This procedure has not yet been subjected to numerical analysis. The task appears to be nontrivial if a relationship with the effect on process limits is to be established, due to the general difficulties in analyzing the (nonlinear) *MCL* process. However, experiments on smaller graphs (with up to thousands of nodes) that allow exact computation indicate that perturbing the process in this manner has very minor impact on the resulting clusterings. The pruning reports in the setting of protein family analysis indicate rather limited pruning of stochastic mass. Additionally, nodes requiring severe pruning can be pruned in advance from the graph to allow for a more precise computation. In this respect, data preprocessing may aid *MCL* the same way it aids approaches to other large scale computational challenges.

Typically for large graphs of several hundreds of thousands of nodes, a maximum K of inbetween $1000-2000$ entries per column is kept. Newly computed columns may contain a number of nonzero entries L amounting to tens of thousands, and selecting the largest K entries from those L using threshold pruning and selection/recovery with min/max heaps has time requirements of order $O(L \log(K))$.

7. Conclusions, further research, and related research. The *MCL* process presented here appears to be both of practical and mathematical interest. A clear relationship was established between *dpsd* matrices, a *DAG* (defined on indices column or rowwise) that can be associated with every such matrix (Theorem 5.4), and the effect of the inflation operator on column stochastic *dpsd* matrices (Theorem 5.5). The *DAG* defined on column indices of *dpsd* matrices generalizes the mapping of nonnegative *MCL*-invariant matrices onto overlapping clusterings, and allows the association of an overlapping clustering with each *dpsd* matrix. In the *MCL* process, the inflation step effectively strengthens the associated *DAG* structure, the expansion step may change it. Many interesting and difficult questions remain. A worthy long standing goal is to prove or disprove Conjecture 1. Two more conjectures are made after the following list of objectives.

(i) For a fixed *MCL* process $(\cdot, e_{(i)}, r_{(i)})$, what can be said about the basins of attraction of the *MCL* process. Are they connected?
(ii) What can be said about the union of all basins of attraction for all limits corresponding with the same overlapping clustering (i.e., differing only in the distribution of attractors)?
(iii) Can the set of limits reachable from a fixed nonnegative matrix M for all *MCL* processes $(M, e_{(i)}, r_{(i)})$ be characterized? Can it be related to a structural property of M?
(iv) Given a node set $I = \{1, \ldots, n\}$ and two directed acyclic graphs D_1 and D_2 defined on I, under what conditions on D_1 and D_2 does there exist a *dpsd* matrix M such that the *DAG*s associated with M according to Theorem 5.4, via rows and columns, respectively, equals D_1 and D_2? What if M is also required to be column stochastic?
(v) Under what conditions do the clusters in the cluster interpretation of the limit of a convergent *MCL* process $(M, e_{(i)}, r_{(i)})$ correspond with connected subgraphs in the associated graph of M?
(vi) For M *dpsd*, in which ways can the *DAG* associated with M^2 be related to the *DAG* associated with M?
(vii) Is it possible to specify a subclass $\mathcal{S}$ of the stochastic *dpsd* matrices and a subset R' of the reals larger than $\mathbb{N}$, such that $\Gamma_r M$ is in $\mathcal{S}$ if $r \in R'$ and $M \in \mathcal{S}$?

Remark. The following is a relaxation of (iv): Given any DAG D is there a symmetric positive semi-definite matrix S such that D is the DAG associated with S (via either columns or rows)? This is easily answered in the affirmative via a constructive and inductive argument, working backwards from sinks to sources, at each step bordering the previously obtained matrix with zeros and adding a suitably constructed rank-one matrix.

Remark. There is no obvious nontrivial hypothesis regarding item (vi), unless such a hypothesis takes quantitative properties of M into account. This is because the breaking up of strongly connected components that can be witnessed in the *MCL* process is always reversible—uncoupling can only happen in the limit. With respect to (v), I conjecture the following.

CONJECTURE 2. *Given a clustering $\mathcal{C}$ associated with a limit of an MCL process with* dpsd *input matrix M, its clusters correspond with subsets of the node set of the associated graph of M that induce connected subgraphs in M.*

Next, consider an *MCL* process $(M, e_{(i)} \stackrel{c}{=} 2, r_{(i)} \stackrel{c}{=} 2)$, with M *dpsd*, that converges towards an *MCL*-invariant matrix L, and let G be the associated graph

of M. The observations in section 2 suggest the following conjecture. Note that a graph automorphism of G implies the existence of a permutation matrix P such that $M = PMP^T$.

CONJECTURE 3. *Each attractor system in L implies that for any pair of elements (k,l) in the attractor system there is a graph automorphism of G mapping k onto l.*

Each instance of two overlapping clusters in L implies the existence of a nontrivial graph automorphism of G, leaving the overlapping part of the two clusters invariant and mapping the remaining part of one of them onto the remaining part of the other.

There are several lines of research that may inspire answers to the questions posed here. However, for none of them the connection seems so strong that existing theorems can immediately be applied. The main challenge is to further develop the framework in which the interplay of Γ_r and Exp_s can be studied. Hadamard-Schur theory was discussed in section 4. Perron-Frobenius theory, graph partitioning by eigenvectors (e.g., [55] and [56]), and work regarding the second largest eigenvalue of a graph (e.g., [1] and [9]), are a natural source of inspiration, and so is the theory of Perron complementation and stochastic complementation as introduced by Meyer ([47] and [48]). There are also papers that address the topic of the structure of matrices which have the subdominant eigenvalue close to the dominant eigenvalue ([30] and [53]). It should be noted that in the former paper matrices are studied that do not have nonnegative spectrum. In the setting of *dpsd* matrices, much stronger results can be expected to hold regarding the relationship between uncoupling measures and spectrum. The literature on the subject of diagonal similarity does not seem to be of immediate further use, as it is often focused on scaling problems (e.g., [17] and [33]). For the study of flip-flop equilibrium states the many results on circulant matrices are likely to be valuable, for example the monograph [10], and the work on group majorization in the setting of circulant matrices in [26]. It may also be fruitful to investigate the relationship with *Hilbert's distance* and the *contraction ratio* for positive matrices, as studied in [4, 6, 8, 27, 58].

Acknowledgments. The author wishes to thank the anonymous referees for their many detailed and useful comments that have significantly contributed to the exposition of the paper.

REFERENCES

[1] N. ALON AND V. D. MILMAN, λ_1, *Isoperimetric inequalities for graphs, and superconcentrators*, J. Combin. Theory Ser. B, 38 (1985), pp. 73–88.

[2] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, Classics in Applied Mathematics 9, SIAM, 1994. Corrected and extended republication of the 1979 book.

[3] R. BHATIA, *Matrix Analysis*, Graduate Texts in Mathematics 169, Springer-Verlag, New York, 1997.

[4] G. BIRKHOFF, *Lattice Theory*, AMS Colloquium Publications 25, 3rd ed., American Mathematical Society, Providence, RI, 1967.

[5] S. BROHÉE AND J. VAN HELDEN, *Evaluation of clustering algorithms for protein-protein interaction networks*, Bioinformatics, 7 (2006) p. 488; available online at http://www.biomedcentral.com/1471-2105/7/488/abstract.

[6] P. J. BUSHELL, *Hilbert's metric and positive contraction mappings in a Banach space*, Arch. Rational Mech. Anal., 52 (1973), pp. 330–338.

[7] Y. CHEN, Y. ZHANG, Y. YIN, G. GAO, S. LI, Y. JIANG, X. GU, AND J. LUO, *Spd—a web-based secreted protein database*, Nucleic Acids Res., 33 (2005), pp. D169–D173.

[8] J. E. COHEN, *Contractive inhomogeneous products of nonnegative matrices*, Math. Proc. Cambridge Philos. Soc., 86 (1979), pp. 351–364.

[9] D. Cvetković and S. Simić, *The second largest eigenvalue of a graph (a survey)*, Filomat, 9 (1995), pp. 449–472.
[10] P. J. Davis, *Circulant Matrices*, John Wiley & Sons, New York, 1979.
[11] S. van Dongen, *A Cluster Algorithm for Graphs*, Technical report, National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam, 2000.
[12] S. van Dongen, *Graph Clustering by Flow Simulation*, Ph.D. thesis, University of Utrecht, The Netherlands, 2000.
[13] B. Dorow and D. Widdows, *Discovering corpus-specific word senses*, in the Tenth Annual Conference of the European Chapter of the Association for Computational Linguistics, Conference Companion, Bergen, Norway, 2003, pp. 79–82.
[14] B. Dorow, D. Widdows, K. Ling, J.-P. Eckmann, D. Sergi, and E. Moses, *Using curvature and Markov clustering in graphs for lexical acquisition and word sense discrimination*, arXiv, (2004) available online at http://arxiv.org/pdf/cond-mat/0403693.
[15] B. Dujon et al., *Genome evolution in yeasts*, Nature, 430 (2004), pp. 35–44.
[16] Eichinger et al., *The genome of the social amoeba dictyostelium discoideum*, Nature, 435 (2005), pp. 43–57.
[17] T. Elfving, *On some methods for entropy maximization and matrix scaling*, Linear Algebra Appl., 34 (1980), pp. 321–339.
[18] A. J. Enright, S. van Dongen, and C. A. Ouzounis, *An efficient algorithm for the large-scale detection of protein families*, Nucleic Acids Res., 7 (2002), pp. 1575–1584.
[19] A. J. Enright, V. Kunin, and C. A. Ouzounis, *Protein families and tribes in genome sequence space*, Nucleic Acids Res., 31 (2003), pp. 4632–4638.
[20] B. S. Everitt, *Cluster Analysis*, 3rd ed., Hodder & Stoughton, London, 1993.
[21] J. Falkner, F. Rendl, and H. Wolkowicz, *A computational study of graph partitioning*, Math. Programming, 66 (1994), pp. 211–239.
[22] M. Fiedler, *Special matrices and their applications in numerical mathematics*, Martinus Nijhoff Publishers, Dordrecht, 1986.
[23] C. H. Fitzgerald and R. A. Horn, *On fractional Hadamard powers of positive definite matrices*, J. Math. Anal. Appl., 61 (1977), pp. 633–342.
[24] A. R. R. Forrest et al., *Phosphoregulators: Protein kinases and protein phosphatases of mouse*, Genome Research, 13 (2003), pp. 1443–1454.
[25] D. Gfeller, J.-C. Chappelier, and P. De Los Rios, *Synonym dictionary improvement through markov clustering and clustering stability*, in Proceedings of the International Symposium on Applied Stochastic Models and Data Analysis, J. Janssen and P. Lenca, eds., 2005, pp. 106–113.
[26] A. Giovagnoli and H. P. Wynn, *Cyclic majorization and smoothing operators*, Linear Algebra Appl., 239 (1996), pp. 215–225.
[27] J. Hajnal, *On products of non–negative matrices*, Math. Proc. Cambridge Philos. Soc., 79 (1976), pp. 521–530.
[28] N. Hall et al., *A comprehensive survey of the plasmodium life cycle by genomic, transcriptomic, and proteomic analyses*, Science, 307 (2005), pp. 82–86.
[29] T. J. Harlow, J. Peter Gogarten, and M. A. Ragan, *A hybrid clustering approach to recognition of protein families in* 114 *microbial genomes*, BMC Bioinformatics, 5 (2004), p. 45.
[30] D. J. Hartfiel and C. D. Meyer, *On the structure of stochastic matrices with a subdominant eigenvalue near* 1, Linear Algebra Appl., 1272 (1998), pp. 193–203.
[31] D. J. Hartfiel and J. W. Speelman, *Diagonal similarity of irreducible matrices to row stochastic matrices*, Pacific J. Math., 40 (1972), pp. 97–99.
[32] D. Heesch and S. Rüger, *NN^k networks for content-based image retrieval*, in McDonald and Tait [45], pp. 253–266.
[33] D. Hershkowitz, W. Huang, M. Neumann, and H. Schneider, *Minimization of norms and the spectral radius of a sum of nonnegative matrices under diagonal equivalence*, Linear Algebra Appl., 241/243 (1996), pp. 431–453.
[34] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 1990.
[35] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, 1991.
[36] N. Jardine and R. Sibson, *Mathematical Taxonomy*, Wiley Series in Probabilistic and Mathematical Statistics, John Wiley & Sons, London, 1971.
[37] V. Kunin, I. Cases, A. J. Enright, V. de Lorenzo, and C. A. Ouzounis, *Myriads of protein families, and still counting*, Genome Biology, 5 (2003), p. 401.

[38] P. Larsson et al., *The complete genome sequence of francisella tularensis, the causative agent of tularemia*, Nature Genetics, 37 (2005), pp. 153–159.
[39] D. A. Lee et al., *Eyesite: A semi-automated database of protein families in the eye*, Nucleic Acids Res., 32 (2004), pp. D148–D152.
[40] R. Leplae, A. Hebrant, S. J. Wodak, and A. Toussaint, *Aclame: A classification of mobile genetic elements*, Nucleic Acids Res., 32 (2004), pp. D45–D49.
[41] L. Li, C. J. Stoeckert, and D. S. Roos, *Orthomcl: Identification of ortholog groups for eukaryotic genomes*, Genome Research, 13 (2003), pp. 2178–2189.
[42] R. Loewy, *Diagonal similarity of matrices*, Portugaliae Mathematica, 43 (1985–1986), pp. 55–59.
[43] L. Lovász, *Random walks on graphs: A survey*, in Miklos et al. [49], pp. 353–397.
[44] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and its Applications*, Mathematics in Science and Engineering 143, Academic Press, New York, 1979.
[45] S. McDonald and J. Tait, eds., *Advances in Information Retrieval: 26th European Conference on IR Research, ECIR* 2004, Lecture Notes in Comput. Sci. 2997, Springer-Verlag, Heidelberg, 2004.
[46] J. McPherson, K.-L. Ma, and M. Ogawa, *Discovering parametric clusters in social small-world graphs*, in The 20th Annual ACM Symposium on Applied Computing, 2005.
[47] C. D. Meyer, *Stochastic complementation, uncoupling Markov chains, and the theory of nearly reducible systems*, SIAM Rev., 31 (1989), pp. 240–272.
[48] C. D. Meyer, *Uncoupling the Perron eigenvector problem*, Linear Algebra Appl., 114/115 (1989), pp. 69–74.
[49] D. Miklos et al., eds., *Combinatorics, Paul Erdős is eighty*, vol. II, Janos Bolyai Mathematical Society, 1996.
[50] H. Minc, *Nonnegative Matrices*, Wiley Interscience Series in Discrete Mathematics and Optimization, John Wiley & Sons, New York, 1988.
[51] B. Mirkin, *Mathematical Classification and Clustering*, Kluwer Academic Publishers, Boston, 1996.
[52] J. Parkinson et al., *A transcriptomic analysis of the phylum nematoda*, Nature Genetics, 36 (2004), pp. 1259–1267.
[53] B. N. Parlett, *Invariant subspaces for tightly clustered eigenvalues of tridiagonals*, BIT, 36 (1996), pp. 542–562.
[54] J. B. Pereira-Leal, A. J. Enright, and C. A. Ouzounis, *Detection of functional modules from protein interaction networks*, PROTEINS: Structure, Function, and Bioinformatics, 54 (2004), pp. 49–57.
[55] A. Pothen, H. D. Simon, and K.-P. Liou, *Partitioning sparse matrices with eigenvectors of graphs*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 430–452.
[56] D. L. Powers, *Structure of a matrix according to its second eigenvector*, in Current trends in matrix theory. Proceedings of the Third Auburn Matrix Theory Conference, Auburn University, Auburn, AL, 1986, F. Uhlig and R. Grone, eds., Elsevier, 1987, pp. 261–265.
[57] B. G. L. Ramaswamy and L. Liu, *Connectivity based node clustering in decentralized peer-to-peer networks*, in the Third International Conference on Peer-to-Peer Computing (ICP2PC 2003), Linkoping, Sweden, 2003.
[58] E. Seneta, *Nonnegative Matrices and Markov Chains*, 2nd ed., Springer, Berlin, 1981.
[59] A. Sinclair, *Algorithms for Random Generation and Counting, A Markov Chain Approach*, Progress in Theoretical Computer Science, Birkhäuser Boston, Boston, 1993.
[60] L. D. Stein et al., *The genome sequence of caenorhabditis briggsae: A platform for comparative genomics*, PLoS Biology, 1 (2003), pp. 166–192.
[61] J. D. Watson et al., *Target selection and determination of function in structural genomics*, IUBMB Life, 55 (2003), pp. 249–255.